Sub-linear Algorithms November 9, 2015

Homework 2

Lecturer: Ronitt Rubinfeld *Due Date: November 23, 2015*

Turn in your solution to each *problem on a separate sheet of paper, with your name on each one.*

1. The goal of this problem is to give a distribution tester with the following behavior: Given access to samples from two distributions p, q both over $[n]$,

   - if $p = q$ output PASS (with probability at least 3/4)
   - if $||p - q||_1 \geq \epsilon$ output FAIL (with probability at least 3/4)

   Let $||p, q|| = \sum p_i q_i$ be the probability that if you pick a sample from p and a sample from q, they are the same element of $[n]$.

   (a) What is $||p - q||_2^2$ in terms of $||p||_2^2, ||q||_2^2, ||p, q||$?

   (b) Given b, an upper bound on the probability of any element of the domain according to p or q. Give an upper bound on the number of samples needed to estimate $||p||_2^2, ||q||_2^2, ||p, q||$ in terms of b.

   (c) Given B, a lower bound on the probability of any element of the domain according to p or q. Give an upper bound on the number of samples needed to estimate $||p - q||_1$ using the naive learning algorithm from lecture 2.

   (d) Set $b = B = \theta(1/n^{2/3})$, devise an algorithm which puts together the previous two parts, using the filtering idea on page 11 of the lecture 3 notes, to get an $\tilde{O}(n^{2/3})$ sample algorithm for the desired distribution tester. Note that you might want to estimate $||p' - q'||_1$ using $||p' - q'||_2^2$ for the distributions p, q restricted to domain elements with probability bounded from above by $b = \theta(n^{2/3})$ (which in turn uses the second part of this question).

2. Given a graph G of max degree d, and a parameter ϵ, give an algorithm for property testing of connectivity. That is, if G is connected, then the algorithm should pass with probability 1, and if G is ϵ-far from connected (at least $\epsilon \cdot dn$ edges must be added to connect G), then the algorithm should fail with probability at least 3/4. Your algorithm should look at a number of edges that is independent of n, and polynomial in d, ϵ. For extra credit, try to make your algorithm as efficient as possible in terms of n, d, ϵ.

   For this homework set, when proving the correctness of your algorithm, it is ok to show that if the input graph G is likely to be passed, then it is ϵ-close to a graph G' which is connected, without requiring that G' has degree at most d.

3. The diameter of an unweighted graph is the maximum distance between any pair of nodes. Give a tester for graphs with degree at most d (where d is a constant and the graph is represented in the adjacency list model) that have low diameter. The tester should have the following specific behavior:

   (a) Graphs with diameter at most D are always accepted.

(b) Graphs which are ϵ-far (that is, at least ϵdn edges must be added) from having diameter $4D + 2$ are failed with probability at least $2/3$.

(c) The query complexity of the tester should be $O(1/\epsilon^c)$ for some constant $1 \leq c \leq \infty$.

For this homework, when proving the correctness of your algorithm, it is ok to show that if the input graph G is likely to be passed, then it is ϵ-close to a graph G' which has diameter $4D + 2$, without requiring that G' has degree at most d.

4. In class we gave an MST approximation algorithm for graphs in which the weights on each edge were integers in the set $\{1..w\}$. Show that one can get an approximation algorithm when the weights can be any value in the range $[1..w]$ (it is ok to get a slightly worse running time).

5. Show a lower bound on giving a multiplicative estimate on the MST: Give two distributions over graphs of degree at most d and weights in the range $\{1, \ldots, w\}$ such that

   (a) graphs in one distribution have an MST weight that is at least twice the MST weight of the graphs in the in other distribution

   (b) in order to distinguish the two distributions with constant probability of success, one must make at least $\Omega(w)$ queries

   If you can get the lower bound to have some nontrivial dependence on d and ϵ, even better!